Statistical hybrid detection and the inference of ancestral distribution areas in *Tolpis* (Asteraceae)

MICHAEL GRUENSTAEUDL1*, BRYAN C. CARSTENS2, ARNOLDO SANTOS-GUERRA3 and ROBERT K. JANSEN4,5

1Institut für Biologie-Botanik, Dahlem Centre of Plant Sciences, Freie Universität Berlin, Altensteinstrasse 6, 14195 Berlin, Germany
2Department of Evolution, Ecology & Organismal Biology, Ohio State University, Columbus, OH 43210, USA
3Calle Guaidil No. 16, CP 38280, Tegueste, Tenerife, Islas Canarias, Spain
4Department of Integrative Biology, University of Texas at Austin, 1 University Station C0930, Austin, TX 78712, USA
5Biotechnology Research Group, Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Received 23 August 2016; revised 28 September 2016; accepted for publication 7 October 2016

Many historical biogeographic studies do not account for the effect of hybrid taxa on phylogenetic tree inference, despite recent advances in the statistical identification of such taxa. This investigation aims to illustrate the impact that hybrid taxa can have on ancestral area reconstructions of the plant genus *Tolpis*, which displays an evolutionary history possibly indicative of a back-colonization of the continental Mediterranean. We evaluate and apply two statistical hybrid detection methods, JML and STEM-hy, which assist in identifying reticulate patterns of allele coalescence. We also evaluate and apply a software tool, P2C2M, to test the fit of genetic loci to the multispecies coalescent model (MSCM). The application of these tools to a previously published data set of three nuclear DNA markers indicates the presence of several potential hybrid taxa in *Tolpis*. One nuclear marker is found to display a reduced level of reticulate history, a good fit to the MSCM and minimal signal of gene flow across archipelagoes. Ancestral distribution areas are reconstructed on gene and species trees of *Tolpis* before and after the exclusion of putative hybrid taxa using stochastic character mapping, parameterized likelihood reconstructions, and reconstructions under continuous-time Markov chain models. The results of these reconstructions indicate that taxa of hybrid origin may have a considerable impact on ancestral area reconstruction and that it is important to account for such taxa prior to biogeographic analysis. We conclude that *Tolpis* has likely had a time-consistent distribution in island habitats and originated on the Canary Islands.


INTRODUCTION

Homoploid hybrid and allopolyploid species can have significant effects on phylogenetic tree inference (Alvarez & Wendel, 2003). In gene tree estimation, DNA sequences of hybrid taxa may not only be recovered in close relation to their respective parents, but also at or near the base of the clade containing both parents, in a position basal to all ingroup taxa, or as sister to an unrelated taxon (McDade, 1992; Soltis et al., 2008). Tree inference under cladistic criteria may be particularly impacted by hybrids formed by distantly related species that are included alongside their parents; their presence in a data set may distort the inference of relationships even for unrelated taxa (McDade, 1990, 1992). Furthermore, the inferred phylogeny may be less resolved than trees estimated without hybrid species (Nieto-Feliner, Fuertes-Aguilar & Rosselló, 2001; Moody & Rieseberg, 2012). Hybrid and allopolyploid species also affect analyses performed subsequently to
tree inference, such as the reconstruction of ancestral character states (see Supporting Information, Fig. S1). If the tree topology in an ancestral character reconstruction displays uncertainty (e.g. due to the presence of hybrid species), the character reconstruction will also be uncertain (Pagel, Meade & Barker, 2004). This dependency necessitates that uncertainty in the phylogeny reconstruction should be accounted for prior to the character reconstruction (Ronquist, 2004).

Two strategies have often been employed to reduce the impact of hybridization among taxa on phylogenetic tree inference and subsequent analyses. Investigators have excluded suspected or known hybrids or allopolyploids from their data sets or removed DNA sequences that displayed reticulate relationships during tree inference (i.e. ‘diploid-only data sets’; Clarkson et al., 2004; Timme, Simpson & Linder, 2007). Alternatively, investigators have utilized direct amplicon DNA sequences instead of isolating allele sequences via cloning despite a priori knowledge of hybrid taxa among the study organisms (e.g. Tippery & Les, 2011; Toepel et al., 2011). However, tree inference based on direct amplicon sequences is unlikely to recover true species relationships because they represent only a subset of the true homolog diversity (Nieto-Feliner & Rosselló, 2007). Both of these strategies are only applicable if the identity of hybrid taxa included in a data set is known a priori.

Recent advances in phylogeny inference and statistical hybrid detection have facilitated an integrated approach towards phylogenetic uncertainty caused by hybrid taxa. For example, the inference of species trees under the multispecies coalescent model (MSCM) can account for reticulations in the underlying gene trees (e.g. Yu et al., 2014) or enable the removal of data sets that do not fit the coalescent model prior to tree inference (e.g. Gruenstaeudl et al., 2016). Also, several methods for the statistical identification of hybridization events have been developed (e.g. Kubatko, 2009; Joly, 2012; Yu, Degnan & Nakhleh, 2012), which can help identify and account for the phylogenetic uncertainty introduced by hybrid or allopolyploid species. However, most of these methods have yet to be evaluated and adopted for historical biogeography, where recent investigations have not explicitly addressed hybrid and allopolyploid taxa (e.g. Zhang et al., 2014; Fougère-Danezan et al., 2015; Xiang et al., 2015). Here, we assess the impact that taxa of hybrid origin can have on ancestral area reconstruction (AAR) and evaluate techniques that may detect their presence.

The plant genus Tolpis Adans. (Asteraceae) is ideally suited to illustrate the impact of hybrid taxa on AAR. Tolpis inhabits four of the five archipelagoes that constitute Macaronesia (Jarvis, 1980) and has presumably colonized each archipelago only once (Gruenstaeudl, Santos-Guerra & Jansen, 2013). Two species are distributed in the continental Mediterranean regions of Europe and North Africa (hereafter ‘mainland’), and the genus comprises at least three species suspected of hybrid origin (Fig. 1): the diploid species *Tolpis crassiuscula* Svent. and *Tolpis succulenta* (Dryand. in Aiton) Lowe and the tetraploid species *Tolpis glabriscens* Kaemmer (Gruenstaeudl et al., 2013). *Tolpis* was also suggested to have colonized the mainland from an island habitat (Moore et al., 2002), a pattern that would place *Tolpis* into a distinct group of Macaronesian plant lineages that have experienced a back-colonization of the continent (Mort et al., 2002; Allan et al., 2004; Carine et al., 2004; Caujapé-Castells, 2011).

We attempt to assess the impact that taxa of hybrid origin have on the biogeographic reconstructions of *Tolpis* by applying three types of analyses to a previously published multigene data set of the genus. First, we infer species trees of *Tolpis* following the identification and exclusion of nuclear loci that do not fit the MSCM. Second, we evaluate and apply statistical methods designed to identify taxa with a reticulate evolutionary history. Third, we reconstruct ancestral distribution areas on gene and species trees with different stochastic and model-based methods. Distribution areas are reconstructed before and after the exclusion of taxa statistically identified as potential hybrids. On the basis of these analyses, we discuss the hypothesis of a back-dispersal to the mainland by *Tolpis*.

MATERIALS AND METHODS

Generalizations on hybrid speciation

Different forms of plant hybridization exist, which result in different types of hybrid species (Soltis & Soltis, 2009; Gompert & Buerkle, 2016). All contain a horizontal exchange of genetic material, which generates a characteristic genetic signal that can be detected and reconstructed as a reticulate pattern of allele coalescence (Linder & Rieseberg, 2004). Most current statistical methods for the detection of such reticulations do not differentiate between the precise forms of hybridization or reliably differentiate between palaeo- and more recent hybridizations. We, therefore, collectively refer to species with reticulate patterns of allele coalescence as ‘hybrid taxa’ or ‘taxa of hybrid origin’, given the understanding that their age and genetic origin may be more diverse than can be discussed here (Gompert & Buerkle, 2016).

DNA sequence data

A previously published nucleotide sequence data set was employed for this investigation (Gruenstaeudl et al., 2013). It comprises DNA sequence alignments...
of two low-copy nuclear markers (A19 and B12, abbre-
viated 'LCNM') and the nuclear ribosomal external
transcribed spacer (ETS) region. The two LCNM are
collectively referred to as locus set 1 (LS1), the set of
all three nuclear DNA markers as locus set 2 (LS2).
Each alignment contains sequences of every recog-
nized Tolpis species (Jarvis, 1980), including a novel
species from the island of La Palma (Crawford, Mort
& Archibald, 2013) and sequences from a potentially
novel species of La Gomera (Gruenstaeudl et al.,
2013). The alignments include sequences from Arnoseris minima (L.) Schweigg. & Koerte, the putative sister taxon to Tolpis (Gruenstaeudl et al., 2013). Each Tolpis spe-
cies is represented by ten allele sequences per locus,
which had been isolated from PCR products via TOPO
TA cloning and then Sanger sequenced (Gruenstaeudl
et al., 2013). Alignment length, number of variable and
parsimony informative sites, maximum divergence of
allele sequences of selected taxa (including those of
potential hybrids), and resampling support for clades
of interest are provided in Table 1.

Figure 1. The maximum likelihood tree of Tolpis as inferred from ETS allele sequences. Circled nodes represent the
MRCAs of the three focal clades. For the purpose of a clearer visualization and only in this figure, allele sequences that stem
from the same species and form exclusive clades were collapsed into single terminals. Alleles of taxa suspected to be hybrids
by previous investigations are underlined. Geographic distributions of extant taxa are indicated by the bars on the right.
Grey bars indicate an exclusive distribution in island habitats. Archipelago abbreviations used: AZ, Azores; CI, Canary
Islands; CM, Continental Mediterranean; CV, Cape Verde; MA, Madeira. Island abbreviations used: C, Gran Canaria; G, La
Gomera; H, El Hierro; P, La Palma; T, Tenerife.

Gene and species tree estimation
To maintain consistent statistical frameworks across
different hybrid detection methods, parallel sets of
software programs were employed for gene and spec-
cies tree estimation. In the Bayesian framework, gene
tree estimation was conducted with BEAST v.1.8
(Drummond et al., 2012) and species tree estimation
Table 1. Alignment length, number of variable sites, number of parsimony informative sites, best-fitting model of nucleotide substitution, largest \( p \)-distance within each taxon set, and phylogenetic resampling support for each clade of the loci under study.

<table>
<thead>
<tr>
<th>Locus or locus set</th>
<th>A19</th>
<th>B12</th>
<th>ETS</th>
<th>LS 1</th>
<th>LS 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of variable sites</td>
<td>133 (30.2%) [109 (25.1%)]</td>
<td>84 (20.0%) [73 (17.4%)]</td>
<td>502 (37.5%) [448 (33.8%)]</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Number of PI sites</td>
<td>86 (19.5%) [84 (19.3%)]</td>
<td>61 (14.5%) [61 (14.6%)]</td>
<td>306 (22.7%) [296 (22.3%)]</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Best-fitting model</td>
<td>HKY + G</td>
<td>TrN + I</td>
<td>GTR + G</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Max. ( p )-distance within <em>Tolpis</em></td>
<td>0.184</td>
<td>0.112</td>
<td>0.192</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Max. ( p )-distance among Canarian endemics of <em>Tolpis</em></td>
<td>0.176</td>
<td>0.087</td>
<td>0.121</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Max. ( p )-distance among clones of <em>T. succulenta</em> from Madeira</td>
<td>0.007</td>
<td>0.073</td>
<td>0.046</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Max. ( p )-distance among clones of <em>T. glabrescens</em></td>
<td>0.119</td>
<td>0.020</td>
<td>0.020</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Max. ( p )-distance among clones of <em>T. crassiuscula</em></td>
<td>0</td>
<td>0</td>
<td>0.018</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
</tbody>
</table>

\( p \)-distances were corrected using the best-fitting nucleotide substitution model and averaged among clone sequences. DNA sequence statistics and \( p \)-distances are not applicable to the marker sets, which were not concatenated. Numbers in parentheses indicate percentages of total alignment length. By default, metrics were calculated on data sets that included the outgroup; metrics calculated on data sets without the outgroup are given in square brackets. bp, base pairs; BS, bootstrap (under maximum likelihood); max., maximum; n.a., not applicable; PI, parsimony-informative; PP, posterior probability.
with *BEAST v.1.8 (Heled & Drummond, 2010). Markov chains were run for 50 M generations, with every 10 000th generation sampled, the first 20% of sampled generations discarded as burn-in, and every fourth generation retained during subsampling. To ensure compatibility with tests of putative hybrid taxa and model fit, a piecewise-constant model of population size change was employed. Independent sampling and convergence of Markov chain parameters was evaluated with Tracer v.1.6 (Rambaut et al., 2014). Maximum clade credibility trees were generated with TreeAnnotator v.1.8 (Drummond et al., 2012). In the likelihood framework, gene tree estimation was conducted with Garli v.2.0.1 (Zwickl, 2006) and species tree estimation with STEM v.2.1.0 (Kubatko, Carstens & Knowles, 2009). A set of 100 independent search replicates was employed, with every 100th generation per run sampled. Maximum clade credibility trees, the best gene, and species trees inferred under maximum likelihood, DNA sequence matrices, and specifications of the distribution areas were submitted to Treebase (http://treebase.org; submission 17247).

**MSCM FIT AND AMONG-ARCHIPELAGO GENE FLOW**

Fit to the MSCM was evaluated via posterior predictive simulation using P2C2M v.0.6 (Gruenstaeudl et al., 2016). Analyses were conducted on each posterior distribution of gene and species trees using 50 simulation replicates per analysis. The coalescent likelihood function ‘lcwt’ was used as summary statistic. Since the relative ability of P2C2M to identify poor fit to the MSCM has yet to be evaluated for cases of hybridization, its statistical power was evaluated using simulations. Here, 20 data sets, each comprising three sequence alignments, were simulated under the same number of species and allele sequences as observed in the empirical data. Effective sample sizes of the loci were estimated with migrate-n v.3.6.6 (Beerli, 2006), average tree depths of the posterior gene tree distributions with Mesquite v.2.75 (Maddison & Maddison, 2011). Species trees were simulated under a Yule model, gene trees with hybrid-Lambda v.0.2 (Zhu et al., 2015), with allele sequences of hybrid taxa shared equally between both parents. Hybridization was only specified in the third gene of each simulated data set. Nucleotide sequence data were simulated with Seq-Gen v.1.3.2 (Rambaut & Grassly, 1997) under the same substitution models and mutation rates as observed in the empirical data.

In addition to the evaluation of fit to the MSCM, we also assessed the level of incomplete lineage sorting (ILS) across and gene flow between different Macaronesian archipelagoes using Bayesian and likelihood-based hypothesis testing (Goldman, Anderson & Rodrigo, 2000). Details of these analyses are presented alongside Table S1 (see Supporting Information).

**DETECTION OF POTENTIAL HYBRID TAXA**

Two statistical methods were employed to detect potential hybrid taxa in *Tolpis*. First, the presence of reticulate patterns of coalescence was evaluated with Bayesian posterior predictive checking using minimum pairwise sequence distances as test statistic (Joly, McLenachan & Lockhart, 2009). The software implementation of this method (JML v.1.0.1; Joly, 2012) identifies species pairs that are more similar than expected under a model of pure ILS and has been applied to several plant lineages (e.g. Jones et al., 2014; Joly, Heenan & Lockhart, 2014). Specifically, JML determines if ILS is a sufficient measure to explain violations of the MSCM based on a distribution of expected pairwise distances across the posterior distribution of species trees. We apply JML on each locus-specific posterior distribution of gene trees, using diploid heredity scalars and a significance level of $P = 0.05$. Second, the presence of reticulate patterns of coalescence was evaluated through model selection among hybrid species trees (Kubatko, 2009). This method can discriminate between ILS and hybridization and has been integrated into the species tree estimation software STEM (as ‘STEM-hy’; Kubatko, 2009). STEM-hy maximizes the likelihood of a species tree given the probability density of gene trees that display topologies and branch lengths consistent with a shared ancestry between a hybrid and a parental taxon. Analyses in STEM-hy were conducted with three independent runs, a cooling rate of 0.005 and locus-specific scaling of branch lengths. The phylogenetic uncertainty of the species tree estimation was accounted for by conducting analyses over the 100 best species trees (Huelsenbeck, Rannala & Masly, 2000).

Prior to applying statistical hybrid detection methods to *Tolpis*, we evaluated their statistical power using several simulated and one empirical data set. The simulated data sets were parameterized to have the same number of species and allele sequences as the empirical data, but differed in the precise coalescence patterns of the alleles, with selected taxa set up as hybrids between two of the other taxa (see Supporting Information, Fig. S2). Simulation settings were identical to those used in the evaluation of the statistical power of P2C2M. The empirical data set constitutes allele sequences of diploid North American roses (Joly & Bruneau, 2006) and had been used to assess the principal applicability of JML (Joly, 2012); it allows a conservative test of hybrid detection, as it displays no or only minimal indications of allele recombination (Joly & Bruneau, 2006; Joly, 2012). It is also similar to the empirical data sets of *Tolpis* by the absence of...
sequences of polyploid taxa. Gene and species tree distributions of the empirical data set were inferred under the same conditions as described in Joly (2012). To reduce calculation time, the species distribution of the empirical data set was subsampled to 100 post-burnin species trees.

**Ancestral Area Reconstruction**

Three types of AAR were conducted to infer the biogeographic history of *Tolpis*: (1) stochastic character mapping, (2) likelihood-based reconstruction with different dispersal rates to and from archipelagoes, and (3) reconstruction under continuous-time Markov chain (CTMC) island models with an instantaneous rate matrix reflecting geographic distance between distribution areas. The reconstructions were conducted over entire tree distributions to accommodate the uncertainty at different levels of the reconstruction process (Ronquist, 2004). For brevity, only the reconstructions of three nested focal clades were recorded (Fig. 1): the clade comprising all extant species of *Tolpis* (clade 1), the clade comprising all Canarian and continental species of the genus (clade 2), and the clade comprising all *Tolpis* species endemic to the Canary Islands (clade 3).

Reconstructing the ancestral distribution areas of the most recent common ancestor (MRCA) of each of these clades is necessary to assess the general biogeographic history of *Tolpis* and discuss the hypothesis of a back-colonization. While included in all analyses, we do not explicitly discuss the Cape Verde species *Tolpis farinulosa* (Webb J.A. Schmidt; as a member of the Canarian clade of *Tolpis* (Gruenstaeudl et al., 2013), any conclusions on the Canarian species of *Tolpis* also apply to this species.

The primary method of AAR employed was stochastic character mapping (Huelsenbeck, Nielsen & Bollback, 2003), which was performed in Mesquite using the wrapper script set WARACS (Gruenstaeudl, 2016). Terminal taxa of the input phylogenies were coded by geographic location, with each Macaronesian archipelago as well as the mainland coded as separate, discrete character states. Stochastic character mapping was performed before and after the exclusion of taxa that were statistically identified as possible hybrids by STEM-hy.

Reconstructions under maximum likelihood (Pagel, 1999) utilized different rates of dispersal to and from Macaronesian archipelagoes. Five dispersal models were evaluated: a symmetrical model in which mainland-to-archipelago dispersal occurred with the same probability as dispersal in the reverse direction (M1), two asymmetrical models with mainland-to-archipelago dispersal set to 50 and 100% less likely than a reverse dispersal (M4 and M5, respectively). Model M1 constitutes a Markov k-state 1 parameter model (Lewis, 2001) and was used to estimate dispersal rates directly. The asymmetrical models (M2–M5) constitute Markov k-state 2 parameter models. Reconstructions were performed on maximum clade credibility trees of the posterior tree distributions and compared on the basis of likelihood scores (Posada & Crandall, 2001).

For reconstructions under CTMC island models, dispersal events were modelled by an instantaneous rate matrix that defined relative distances between the distribution areas of *Tolpis* as transition probabilities (Ronquist & Sanmartin, 2011). Optimal reconstructions were inferred via MCMC sampling in MrBayes v.3.2.3 (Ronquist et al., 2012) in a combined analysis of sequence data and geographic characters. Topology-constrained Markov chains were run for 20 M generations, with 50% discarded as burnin. The geographic data partition was analysed under a general time-reversible model, the DNA sequence partitions under the best-fitting models of nucleotide substitution. Dispersal probabilities were modelled under a negative exponential distribution. Dispersal rate and nucleotide sequence evolution were estimated independently with unlinked rate priors across partitions.

**RESULTS**

**MSCM Fit and Among-Archipelago Gene Flow**

Simulation testing using P2C2M indicates that it can accurately detect cases of poor model fit caused by hybridization with a significance level of $P = 0.10$. However, this low rate of type II error came at the price of a high rate of type I error (Fig. 2). The selection of more stringent significance levels decreased the rate of type I and increased the rate of type II error. Under $P = 0.01$, only one out of 40 genes simulated under the MSCM exhibited poor model fit, while three genes simulated under hybridization fit the model. The application of P2C2M on the empirical data of *Tolpis* identified poor model fit for the ETS at $P = 0.01$ (Table 2). In the estimation of LS2 species trees, it was the only locus to not fit the MSCM. The LCNM, by contrast, were identified to fit the coalescent model both under LS1 and LS2 species trees.

Support for the presence of ILS across or gene flow between different archipelagoes in *Tolpis* was identified under both Bayesian and likelihood-based hypothesis testing (see Supporting Information, Table S1). The results of the AU tests showed that trees of each posterior gene tree distribution became significantly worse when they were topologically constrained to exclude among-archipelago ILS or gene flow. The same result was recovered by log Bayes factors. Species trees based
on all three loci also displayed worse likelihoods and log Bayes factors when constrained. LS1 species trees, however, displayed a level of ILS or gene flow that was not distinguishable from homoplasy. Here, the AU test did not recover a significant difference between constrained and unconstrained trees. Moreover, the log Bayes factor for this comparison was found positive, indicating support for very low levels of ILS across or gene flow between archipelagoes. In summary, the posterior distribution of LS1 species trees was least affected by a topological constraint, implying the absence of among-archipelago ILS or gene flow. We thus considered AAR conducted on the LS1 posterior tree distribution to be more reliable than AAR on the other tree distributions.

**EVALUATION OF STATISTICAL HYBRID DETECTION**

Both hybrid detection methods indicated the presence of taxa with reticulate evolutionary histories in the simulated data sets. However, only STEM-hy exhibited an acceptable statistical specificity and sensitivity. STEM-hy, which conducts tests for hybridization under all loci collectively, correctly identified the involvement of the hybrid and one parental taxon in more than three quarters of the simulated data sets (see Supporting Information, Fig. S3). Incorrect inferences of hybridization existed, but occurred in less than one quarter of the simulations. JML, which conducts tests for hybridization in each locus independently, displayed lower specificity and sensitivity. It displayed high rates of type I error by indicating hybridization where none was modelled (17/20 under $P = 0.05$; 10/20 under $P = 0.01$) and simultaneously displayed medium rates of type II error by identifying modelled hybridization in less than half of the simulations (9/20 under $P = 0.05$; 6/20 under $P = 0.01$; see Supporting Information, Fig. S4). A concurrent result was obtained when comparing both methods and P2C2M under the empirical data of Joly & Bruneau (2006). While P2C2M identified poor model fit to the MSCM under the very gene that also indicated presence of hybridization with JML (see Supporting Information, Table S2, Fig. S5a; Joly, 2012), STEM-hy did not recover indications for hybridization in this data set (except for cases of false positives; see Supporting Information, Fig. S5b). None of the potential hybridizations inferred with JML by Joly (2012) was statistically significant at $P = 0.05$, which renders the presence of hybrids among the analysed *Rosa* species unlikely and supports the results inferred by STEM-hy. In summary, STEM-hy appears to generate more conservative inferences of hybrid detection than JML. For the present investigation, we therefore selected STEM-hy for the detection of potential hybrid taxa and, by extension, their exclusion from the data sets for AAR.
Potential hybrid taxa in *Tolpis*

Under the empirical data of *Tolpis*, STEM-hy inferred the participation of *Tolpis macrorhiza* (Lowe ex Hook.) DC. in hybridizations under every species tree and *Tolpis barbata* (L.) Gaertn. in more than three quarters of the species trees (Fig. 3). Other taxa are identified for hybridization in less than one fifth of the species trees, which we interpreted as false positives given the results of the corresponding simulations. JML, by comparison, indicated numerous cases of hybridization as well as a conspicuous correlation between the number of inferred hybridizations and the amount of genetic divergence per locus (see Supporting Information, Fig. S6). Under locus A19 and the ETS, hybridization was inferred among the majority of Canarian endemics as well as between the Canarian and the mainland species. The widely distributed continental and insular species *T. barbata* was inferred to partake in many of these hybridizations. Under locus B12, only a single instance of hybridization was inferred: between *T. glabrescens* and *T. succulenta*. The same hybridization event was inferred under locus A19. For the purpose of evaluating the effect on AAR, we followed the results of STEM-hy and excluded the alleles of *T. barbata* and *T. macrorhiza* from our DNA alignments and recalculated the gene and species trees.

Ancestral area reconstruction

Ancestral area reconstruction by stochastic character mapping conducted prior to the exclusion of potential hybrid taxa recovered different results for different posterior tree distributions (Fig. 4; see Supporting Information, Figs S7 and S8). For clade 1, the Canary Islands were identified as the most probable ancestral distribution area under locus B12, Madeira under locus A19 as well as LS1 species trees, and the mainland under the ETS as well as LS2 species trees. For clade 2, the Azores or the Canary Islands were identified as ancestral distribution area under locus A19, the Canary Islands alone under LS1 species trees, and the mainland under locus B12, the ETS, and LS2 species trees. Ancestral area reconstruction conducted after the exclusion of *T. barbata* and *T. macrorhiza* was more homogeneous (Fig. 4). The reconstruction results indicated that each of the current habitats was a potential ancestral distribution area for clade 1, with exception of the Canary Islands under the ETS. For clade 2, the Canary Islands were inferred the most probable ancestral distribution area under each locus or locus set, except under the ETS. The reconstructions for clade 3 were consistent across different posterior tree distributions before and after the exclusion of *T. barbata* and *T. macrorhiza* in inferring the Canary Islands as the ancestral distribution area. AAR under the LS1 species trees favoured the Canary Islands as the ancestral distribution for each of these clades.

Ancestral area reconstructions under maximum likelihood and parameterized dispersal models suggested a biogeographic history of *Tolpis* that favoured dispersal from continental to island habitats. Models M2 and M3 produced the best likelihood scores in all reconstructions, except for the gene tree of B12, for which model M1 was suggested to be most likely given the sequence data (Table 3). Reconstructions under models favouring a back-dispersal to the mainland.
were found to fit the data worse than mainland-to-island or symmetric reconstruction models.

Ancestral area reconstructions under CTMC island models indicated support for a Canarian origin of the genus under loci A19 and B12 and, conversely, support for a continental origin of the genus under the ETS (Table 4). Likewise, the ancestral distribution area of the MRCA of all Canarian and mainland species of *Tolpis* was reconstructed to the Canary Islands under loci A19 and B12 and to the mainland under the ETS. The same results were recovered for clade 3.

### DISCUSSION

**APPLICATIONS OF STATISTICAL HYBRID DETECTION**

Taxa of hybrid origin can have a considerable impact on phylogenetic tree inference, which should be
considered as a source of error for cladistic analysis (McDade, 1992; Soltis et al., 2008). This impact is illustrated by the multitude of phylogenetic positions that hybrid taxa can be recovered in (McDade, 1990, 1992), a phenomenon that is ultimately caused by different levels of sequence similarity and homoplasy between hybrid taxa and their parents due to the age and type of hybridization (Nieto-Feliner et al., 2001; Soltis et al., 2008). In particular, the inclusion of hybrid taxa formed by distantly related parents in a data set that also comprises these parents may lead to spurious phylogenetic placements, including the recovery of hybrid taxa in a clade with unrelated species (McDade, 1992). The identification of putative hybrid taxa is thus an important first step in counteracting their potential impact on phylogenetic tree inference and subsequent analyses.

In order to detect and, where necessary, exclude taxa of potential hybrid origin from our data sets prior to AAR, we evaluated and applied two statistical methods to detect reticulate patterns of allele coalescence as well as one statistical method to evaluate the fit of genetic loci to the MSCM. Each of these methods operates on DNA sequence data and evaluates a specific aspect commonly associated with hybrid speciation (Linder & Rieseberg, 2004). The methods display several advantages compared to more traditional evaluations of hybrid speciation such as artificial cross-pollination or cytogenetic experiments (Chester et al., 2010). First, statistical hybrid detection is generally less expensive and less time-consuming than more traditional approaches, requiring only DNA sequence data of orthologous genomic regions. Second, statistical hybrid detection can also be applied to those hybrids that cannot be introgressed to potential parents due to hybrid sterility or the onset of parthenogenesis. Third, statistical hybrid detection is the only applicable method for investigations that rely on desiccated plant material. Thus, statistical hybrid detection is a vital tool for a wide range of investigations. However, as any statistical test, these methods contain an intrinsic statistical error that needs to be evaluated and taken into account when interpreting the results.

### Table 4. Results of reconstructions under CTMC island models with an instantaneous rate matrix reflecting geographic distance between archipelagoes

<table>
<thead>
<tr>
<th>Locus</th>
<th>Archipelago</th>
<th>Clade 1</th>
<th>Clade 2</th>
<th>Clade 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene tree – A19</td>
<td>Azores</td>
<td>0.006 (±0.000)</td>
<td>0.000 (±0.000)</td>
<td>0.000 (±0.000)</td>
</tr>
<tr>
<td></td>
<td>Continent</td>
<td>0.336 (±0.138)</td>
<td>0.000 (±0.000)</td>
<td>0.039 (±0.008)</td>
</tr>
<tr>
<td></td>
<td>Madeira</td>
<td>0.124 (±0.007)</td>
<td>0.032 (±0.004)</td>
<td>0.026 (±0.004)</td>
</tr>
<tr>
<td></td>
<td>Canary Islands</td>
<td>0.534 (±0.101)*</td>
<td>0.968 (±0.004)*</td>
<td>0.935 (±0.010)*</td>
</tr>
<tr>
<td>Gene tree – B12</td>
<td>Azores</td>
<td>0.000 (±0.000)</td>
<td>0.000 (±0.000)</td>
<td>0.000 (±0.000)</td>
</tr>
<tr>
<td></td>
<td>Continent</td>
<td>0.002 (±0.000)</td>
<td>0.043 (±0.023)</td>
<td>0.001 (±0.000)</td>
</tr>
<tr>
<td></td>
<td>Madeira</td>
<td>0.083 (0.065)</td>
<td>0.026 (±0.009)</td>
<td>0.000 (±0.000)</td>
</tr>
<tr>
<td></td>
<td>Canary Islands</td>
<td>0.915 (±0.065)*</td>
<td>0.932 (±0.039)*</td>
<td>0.998 (±0.000)*</td>
</tr>
<tr>
<td>Gene tree – ETS</td>
<td>Azores</td>
<td>0.003 (±0.001)</td>
<td>0.000 (±0.000)</td>
<td>0.000 (±0.000)</td>
</tr>
<tr>
<td></td>
<td>Continent</td>
<td>0.659 (±0.188)*</td>
<td>0.980 (±0.000)*</td>
<td>0.946 (±0.002)*</td>
</tr>
<tr>
<td></td>
<td>Madeira</td>
<td>0.336 (±0.187)</td>
<td>0.007 (±0.000)</td>
<td>0.001 (±0.000)</td>
</tr>
<tr>
<td></td>
<td>Canary Islands</td>
<td>0.003 (±0.000)</td>
<td>0.013 (±0.000)</td>
<td>0.053 (±0.002)</td>
</tr>
</tbody>
</table>

The variances of the reconstruction results are given in parentheses. Asterisks indicate the reconstructions with the highest posterior state probabilities.

### Evaluation of statistical hybrid detection

In this study, STEM-hy was found to be a more conservative hybrid detection method with preferable statistical power than JML (see Supporting Information, Figs S3–S5). Similar results have been reported by Heled, Bryant & Drummond (2013), who indicated that JML might suffer from elevated rates of type I error. However, our findings on the relative error rates of STEM-hy and JML are contingent on the precise dimensions of the data evaluated and do not invalidate JML as a whole. More investigation is needed to estimate the sensitivity and specificity of both methods, particularly regarding the level of genetic divergence necessary to differentiate ILS from hybridization. In particular, the results of JML point to a positive correlation between the number of reticulations detected and the amount of variation of each locus as well as the amount of genetic divergence among them. For example, the ETS has the highest number of variable sites and the largest maximum p-distance between ingroup taxa (Table 1) while displaying the largest number of inferred hybridization events under JML. The LCNM B12, by contrast, has the smallest number of variable sites and the smallest maximum p-distance between ingroup taxa, and it also displays the smallest number
of inferred hybridization events. It could, therefore, be argued that the ability of JML to differentiate ILS from hybridization may be sensitive to the overall level of genetic divergence among the input sequences. In summary, our evaluations do not provide proof for or against the integrity of either method, but allow a first estimation of their relative error rates and provide a model for future analyses.

**INDICATIONS FOR HYBRIDIZATION IN *TOLPIS***

STEM-hy indicated that the continental and insular taxon *T. barbata* as well as the Madeiran endemic *T. macrorhiza* were participating in hybridization events (Fig. 3). The results of STEM-hy are partially consistent with those of JML, which indicated the participation of *T. barbata* in hybridizations in two loci (see Supporting Information, Fig. S6). As the only widespread and potentially invasive species of the genus, *T. barbata* displays characteristics indicative of many hybrid taxa (Rieseberg et al., 2007). Similarly, *T. macrorhiza* is the only Macaronesian species of *Tolpis* that exhibits a low proportion of flowering individuals among natural populations in any given year (Crawford et al., 2015); a potential connection between hybridization and floral development has been reported for Asteraceae (e.g. Bello et al., 2013). The hybridization events suggested by STEM-hy are consistent with the findings of Soto-Trejo et al. (2013), who report that members of the Canarian clade of *Tolpis* can be readily hybridized in cultivation. For the purpose of evaluating the effect on AAR, *T. barbata* and *T. macrorhiza* were thus excluded from the empirical data sets during AAR of *Tolpis*. However, additional experimental work is warranted to confirm the true hybrid nature of both species.

The three species of *Tolpis* suggested to be hybrids by a previous molecular investigation (i.e. *T. crassiuscula*, *T. succulenta*, and *T. glabrescens*; Gruenstaeudl et al., 2013) were not inferred as potential hybrids by STEM-hy. Speculation of hybrid identity in *Gruenstaeudl* et al. (2013) was largely based on the observation that species that do not inhabit the same island or even the same archipelago shared identical or highly similar ETS alleles. To conclude hybridization from this observation may have been premature, as several factors can lead to shared alleles across allopatric species. For example, nuclear ribosomal DNA is prone to concerted evolution, which streamlines allele diversity across species (Alvarez & Wendel, 2003; Poczai & Hyvönen, 2010). Likewise, ILS in recently diverged species may group together allele sequences of nonsister species during gene tree estimation (Linder & Rieseberg, 2004; Joly et al., 2009). False inferences of hybridization can also occur when introgressed alleles stem from multiple, closely related sources because the independent introduction of identical alleles can obscure their inference as shared or derived (Eaton et al., 2015). In summary, simply observing shared alleles across nonsister species upon gene tree estimation is insufficient to conclude that hybridization has occurred. At a minimum, a statistical evaluation to determine if the observed relationships can be explained by ILS alone is necessary (Joly, 2012).

**MSCM FIT AND AMONG-ARCHIPELAGO GENE FLOW**

When discordance between different genealogical histories exists, character reconstructions performed on species trees should allow more consistent conclusions than those on individual gene trees. However, not all genetic loci necessarily fit the MSCM, even though this is the basic assumption of most species tree estimation methods (Reid et al., 2014). The identification and exclusion of loci with poor fit to the MSCM is an important prerequisite for the estimation of species trees (Kubatko, 2009) and subsequent AAR. P2C2M was developed to help identify loci with poor fit to the MSCM (Gruenstaeudl et al., 2016). Prior to the empirical assessment of model fit, we conducted a set of simulations to assess the ability of P2C2M to detect cases in which poor fit is caused by hybridization. The results of these simulations indicated that P2C2M had a high statistical power when poor model fit was caused by hybridization (Fig. 2) and that it can be utilized in the identification of hybridizations under stringent significance levels.

In order to identify and, where necessary, exclude genes from the data sets prior to AAR, we utilized P2C2M on the empirical data of *Tolpis*. P2C2M indicated that both LCNM fit the MSCM (Table 2). Hence, we decided to consider only those AAR results that were generated on LS1 species trees. DNA sequences of ETS alleles, by contrast, were not found to fit the MSCM, which coincides with the observation that they display the highest relative number of variable sites and the greatest p-distance among the loci employed (Table 1). The high genetic variability of the ETS may be indicative of the elevated rates of non-homologous recombination observed in nuclear ribosomal DNA (Poczai & Hyvönen, 2010). Concerted evolution in nuclear ribosomal loci, which streamlines intraspecific allele diversity and can erase hereditary information, is often incomplete, generating novel or rearranged sequence homologs, which can facilitate non-homologous recombination (Alvarez & Wendel, 2003). It is not implausible that the lack of fit to the MSCM by the ETS is partially caused by nonhomologous recombination.

As part of this investigation, we also assessed the level of ILS across and gene flow between
different Macaronesian archipelagos (see Supporting Information, Table S1). Both processes can be detrimental to the reconstruction of ancestral character states and warrant pre-analysis identification. The results indicated that the posterior distribution of LS1 species trees displayed the least amount of ILS or gene flow across archipelagoes, which corroborated our selection of LS1 species trees as the basis of AAR in Tolpis. Support for the scenario of reduced gene flow across archipelagoes comes from Borges-Silva et al. (2016), who highlight the sea as a geographical barrier to gene flow in Tolpis.

**Ancestral Area Reconstruction**

Our tests for hybridization, ILS, and among-archipelago gene flow highlight the need for selection among the AAR results. Thus, we only considered reconstructions that were generated (1) on data with minimal signal for hybridization, ILS, or poor model fit, and (2) after the exclusion of potential hybrid taxa. In addition, we preferred reconstructions conducted on species trees over those on gene trees in order to avoid bias through differential allele divergence. Ancestral area reconstruction on LS1 species trees inferred without *T. barbata* and *T. macrocarpa* conformed to these prerequisites. The results identified the Canary Islands as the most probable ancestral distribution area for the genus as a whole (Fig. 4), a finding supported by the observation that the majority of extant *Tolpis* species occur on this archipelago. The results under post-exclusion LS1 species trees also support the Canary Islands as the most probable ancestral distribution area for the continental taxa of *Tolpis*. The reconstructions under CTMC island models corroborate both conclusions (Table 4). Given these results, it seems likely that *Tolpis* has consistently inhabited island habitats and undergone one or more dispersals from an island to a continental habitat during the course of its evolutionary history. Assuming a continental origin of the genus in the distant past, *Tolpis* underwent at least five separate dispersal events to reach its extant distribution (Fig. 5), which is the same number of dispersal steps inferred by Moore et al. (2002).

The application of likelihood-based reconstructions under parameterized models of dispersal inferred a biogeographic history that favoured dispersal following the currents of the north-eastern trade winds. These winds are prevalent throughout the Macaronesian region and are thought to have mediated the dispersal of many plant species (Francisco-Ortega, Jansen & Santos-Guerra, 1996). However, these results should not be taken as evidence against a back-colonization because the log likelihoods estimated under different dispersal models pertain to entire trees, not specific clades. Local deviations from the best-fitting models of dispersal are still possible. Hence, these results could be interpreted as support for a scenario in which the back-dispersal occurred between the Canary Islands and North Africa (Caujapé-Castells, 2011), which at the closest location are only 95 km apart.

**Back-colonization of continent**

The hypothesis of a back-colonization of the continent by *Tolpis* was first suggested by Moore et al. (2002), who identified a re-colonization of the mainland following the extinction of *Tolpis* in Europe and North Africa as the most parsimonious solution. No indication of a back-dispersal by *Tolpis* was found by Gruenstaeudl et al. (2013), but their biogeographic analyses may have been compromised by hybrid taxa. The present investigation provides support for a back-colonization of *Tolpis* by inferring that the MRCA of all extant species has likely inhabited the Canary Islands, necessitating dispersal to the continent to match the current distribution of the genus (Fig. 5). This phenomenon, termed ‘biodiversity boomerang’, has been reported from several Macaronesian plant lineages (Mort et al., 2002; Allan et al., 2004; Carine et al., 2004; Caujapé-Castells, 2011). Biological support for such a scenario is provided by Bellemain & Ricklefs (2008), who concluded that island taxa should be capable of back-dispersal if their adaptation to island life has not resulted in reduced dispersal ability. Although dispersal in plants is predominantly unidirectional (Emerson, 2002), a back-colonization of the continent would remain consistent with this logic due to the close geographic proximity of source and sink area. In fact, the species *T. barbata* inhabits the Canary Islands and most of North Africa and Mediterranean Europe, indicating that repeated dispersal may be common in this lineage.

Despite the support for a back-colonization of the continent by *Tolpis*, this scenario should be treated with caution. Like all character state reconstructions, AAR is sensitive to the precise taxon set under study. The exclusion of two species from the final data set altered the reconstruction results, although it did not alter the inference of a back-colonization (Fig. 4). When the potential hybrid taxa remain in the data set, Madeira is inferred the most probable ancestral distribution area for clade 1 and the Canary Islands for clade 2. The inference of a reverse colonization by *Tolpis* thus seems to be supported irrespective of the exclusion of potential hybrid taxa under LS1, but the inferred colonization pathway changes. Caution about the scenario of a back-colonization is also warranted given the possibility of past extinctions in *Tolpis*. Several investigations have reported likely extinctions in Macaronesian plant lineages, correlating them to climatic fluctuations in the Mediterranean during the Pliocene (e.g.
Figure 5. Phylogenetic relationships in *Tolpis* after the exclusion of potential hybrid taxa and the association with extant distribution areas. The location of the study area in a global context is displayed on the top, a detailed map that displays coastlines and country borders between the 26th meridian west, the 1st meridian east, the 43rd parallel north, and the 16th parallel north in the centre, the maximum clade credibility tree of the posterior species tree distribution inferred under locus set LS1 after the exclusion of potential hybrid taxa at the bottom of the figure. Species–area associations are visualized as dashed gray lines. Arrows next to the branches of the phylogenetic tree indicate dispersal events as inferred through AAR and the destination of each dispersal. The remaining mainland species and the outgroup were arbitrarily connected to the Iberian peninsula, given the understanding that both taxa have a much wider continental distribution. Node numbers on the phylogenetic tree indicate posterior probability values greater than 0.5. The geographic locations of entries of the keyword ‘*Tolpis*’ in the database GBIF (http://www.gbif.org/) within the centre map are displayed as gray dots. *Tolpis* also inhabits the easternmost Canarian islands of Fuerteventura and Lanzarote, but no GBIF records for *Tolpis* on these islands seem to exist; the same is true for the Cape Verde archipelago. Cont., Continental Mediterranean; Isl., Islands.
Vitales et al., 2014). Historical extinctions and incomplete sampling of extant taxa can be important obstacles to reconstructing colonization, as they can mislead the correct interpretation of pathways (Emerson, 2002). Although this investigation employs a comprehensive taxon set that includes all known extant species of Tolpis, and the diversification of Macaronesian plant lineages with similar distributions seems to have been very recent (Kim et al., 2008), the existence of populations on islands where they have since become extinct cannot be discounted. Such historical populations would be particularly likely on the easternmost Canarian islands Fuerteventura and Lanzarote, which are in immediate vicinity of the North African mainland and currently inhabited by Tolpis only through T. barbata. The only moderate support for a clade of all Canarian and continental taxa of Tolpis (Fig. 5) could be indicative of such unsampled alleles.

**Scenarios of among-archipelago gene flow**

The results of the present investigation enable us to infer a possible scenario for among-archipelago gene flow in Macaronesia. Most insular plant lineages colonize specific archipelagoes only once, followed by adaptive radiations on different islands into different habitats (Emerson, 2002). Consequently, haplotypes from species endemic to specific archipelagoes are rarely shared across Macaronesian archipelagoes, but when it happens they are often interpreted as indications for among-archipelago gene flow (Pelser et al., 2012; Talavera et al., 2013). Reports of plant migration between Macaronesian archipelagoes (e.g. Carine et al., 2004) indicate a possible scenario for such gene flow. Allopatric distributions, which constitute the main obstacle for among-archipelago gene flow, can become sympatric distributions through occasional migration, enabling introgression into the gene pool of the local endemics (Pelser et al., 2012). Similarly, gene flow may be mediated by local hybridization and concluded by the local extinction of one parent. The observation that both potential hybrids identified through STEM-hy were found among the older lineages of Tolpis and not among the recently radiated Canarian endemics supports this scenario. Recent progress on a better resolved nuclear phylogeny of Tolpis (Mort et al., 2015) should allow a more detailed evaluation of this scenario.

**Outlook and conclusions**

Next-generation sequencing and improved species tree inference methods will likely advance the statistical detection of hybrid taxa and the effort to minimize their impact on phylogeny inference and subsequent analyses. For example, the sequencing of hundreds of nuclear loci through next-generation sequencing may make it easier to exclude taxa from investigations without greatly compromising the inference of species trees (Streicher, Schulte & Wiens, 2016). Similarly, the use of phylogenomic data sets can considerably improve the inference of species trees compared to Sanger-sequenced data sets (Ruane et al., 2015), which, in turn, can improve the statistical detection of potential hybrid taxa.

This investigation found that taxa of hybrid origin may have a considerable impact on ancestral area reconstruction. We advocate the application of methods that assist in the statistical detection and, where necessary, exclusion of taxa or loci with reticulate evolutionary histories prior to the estimation of species trees and the reconstruction of ancestral distribution areas. However, given that all data sets analysed here comprise only homoploid hybrid species (except for the tetraploid T. glabrescens), we do not recommend generalizing our conclusions to data sets of primarily allopolyploid taxa. Given significant differences in the genetic divergence of the parental species, the number and genetic diversity of alleles in allopolyploids may be substantially different from those of homoploid hybrids (Paun et al., 2009). Additional research is needed to corroborate the hybrid nature of the taxa indicated here and to evaluate the statistical power of the applied methods on allopolyploid taxa. Based on reconstructions adjusted for putative hybrid taxa, we conclude that the plant genus Tolpis has likely had a time-consistent distribution in island habitats and that it likely originated on the Canary Islands. Our results also support the scenario that Tolpis is part of a distinct group of Macaronesian plant lineages that have experienced a back-dispersal to the continent.

**Acknowledgments**

This investigation was supported by the Blake Centennial Professorship in Systematic Botany of The University of Texas at Austin to R.K.J., by a start-up package from The Ohio State University to B.C.C., graduate student research awards from the American Society of Plant Taxonomists and the Botanical Society of America to M.G. and the Plant Resources Center at The University of Texas at Austin. We also thank three anonymous reviewers for their helpful comments. This work was supported in part by an allocation of computer resources from the HPC Service of ZEDAT of the Freie Universität Berlin and from the Ohio Supercomputer Center.

**References**


Heled J, Bryant D, Drummond AJ. 2013. Simulating gene trees under the multispecies coalescent and time-dependent migration. BMC Evolutionary Biology 13:44.


**SUPPORTING INFORMATION**

Additional Supporting Information may be found online in the supporting information tab for this article:

**Figure S1.** Example of the impact of hybrid or allopolyploid taxa on the reconstruction of ancestral character states.

**Figure S2.** Two examples of gene and species trees inferred from simulated data as employed to evaluate the statistical power of STEM-hy and JML in detecting potential cases of hybridization.

**Figure S3.** Results of simulation analyses to evaluate the statistical power of STEM-Hy to detect cases of hybridization among a set of taxa.

**Figure S4.** Results of simulation analyses to evaluate the statistical power of JML to detect cases of hybridization among taxa of specific loci.

**Figure S5.** Results of the evaluation and comparison of the statistical power of JML and STEM-hy when applied to the empirical data set of Joly & Bruneau (2006).

**Figure S6.** Results of statistical hybrid identification with JML on the empirical data sets of Tolpis.

**Figure S7.** Results of AAR prior to the exclusion of Tolpis species that were indicated to be potential hybrids by STEM-hy.

**Figure S8.** Results of AAR after the exclusion of Tolpis species that were indicated to be potential hybrids by STEM-hy.

**Table S1.** Results of an evaluation of the level of incomplete lineage sorting and among-archipelago gene flow in Tolpis via Bayesian and likelihood-based hypothesis testing.

**Table S2.** Results of analyses with P2C2M on the empirical data set of Joly & Bruneau (2006).